Classifying the form of iconic hand gestures from the linguistic categorization of co-occurring verbs

Magdalena Lis
Centre for Language Technology
University of Copenhagen
Njalsgade 140
2300 Copenhagen
magdalena@hum.ku.dk

Costanza Navarretta
Centre for Language Technology
University of Copenhagen
Njalsgade 140
2300 Copenhagen
costanza@hum.ku.dk

Abstract

This paper deals with the relation between speech and form of co-occurring iconic hand gestures. It focuses on multimodal expression of eventualities. We investigate to what extent it is possible to automatically classify gestural features from the categorization of verbs in a wordnet. We do so by applying supervised machine learning to an annotated multimodal corpus. The annotations describe form features of gestures. They also contain information about the type of eventuality, verb Aktionsart and Aspect, which were extracted from plWordNet 2.0. Our results confirm the hypothesis that the Eventuality Type and Aktionsart are related to the form of gestures. They also indicate that it is possible to some extent to classify certain form characteristics of gesture from the linguistic categorization of their lexical affiliates. We also identify the gestural form features which are most strongly correlated to the Viewpoint adopted in gesture.

Keywords: multimodal eventuality expression, iconic co-speech gesture, wordnet, machine learning

1 Introduction

In face-to-face interaction humans communicate by means of speech as well as co-verbal gestures, i.e. spontaneous and meaningful hand movements semantically integrated with concurrent spoken utterances (Kendon, 2004; McNeill, 1992). Gestures which depict entities are called iconic gestures. Such gestures are co-expressive with speech, but not redundant. According to inter alia McNeill (1992; 2005) and Kendon (2004), they form an integral part of a spoken utterance.

Iconic gestures are especially well-suited to express spatio-motoric information (Alibali et al., 2001; Krauss et al., 2000; Rauscher et al., 1996) and, thus, often accompany verbal expressions of eventualities, in particular motion eventualities. Eventuality is an umbrella term for entities like events, actions, states, processes, etc. (Ramchard, 2005).[1] On the level of language, such entities are mostly denoted by verbs. Gesturally, they are depicted by means of iconicity relation (McNeill, 1992; McNeill, 2005; Peirce, 1931). This relation does not, however, on its own fully explain the form that a gesture takes - a referent can be depicted in gestures in multiple ways, for instance from different perspectives - that of the observer or that of the agent. How a speaker chooses to represent a referent gesturally determines which physical form a gesture takes. Knowledge about the factors influencing this choice, is still sparse (Kopp et al., 2008). It is, however, crucial not only for our understanding of human communication but also for theoretical models of gesture production and its interaction with speech. Such models can in turn inform generation of natural communicative behaviors in Embodied Conversational Agents (Kopp et al., 2008).

The present paper contributes to this understanding. It addresses a particular aspect of gesture production and its relationship to speech with focus on multimodal expression of eventualities. Various factors have been suggested to influence eventuality gestures, including referent characteristics (Parrill, 2010; Poggi, 2008), verb Aspect (Duncan, 2002) and Aktionsart (Becker et al., 2011). We present a pilot study investigating the extent to which hand gestures can be automatically

[1] In gesture studies the terms 'action' or 'event' are habitually used in this sense. We adopted the term 'eventuality' to accommodate the terminology for Aktionsart categories reported in Subsection 3.2.2, where 'action' and 'event' are subcategories of what can be termed 'eventualities.'

classified from the information about these factors. We extract this information from the categorization of verbs in a lexical-semantic database called wordnet. The theoretical background and methodological framework are discussed in (Lis, 2012a; Lis, 2012b; Lis, submitted).

In the present paper, differing from preceding studies on the multimodal expression of eventualities, we test the hypotheses by applying supervised learning on the data. Our aim in employing this method is to test the annotation scheme and potential application of the annotations in automatic systems and to study the relationship between speech and gesture not only for relations between single variables but also groups of attributes. In this, we follow the approach adopted by a number of researchers. For example, Jokinen and colleagues (2008) have used classification experiments to test the adequacy of the annotation categories for the studied phenomenon. Louwerse and colleagues (2006a; 2006b) have applied machine learning algorithms on annotated English map-task dialogues to study the relation between facial expressions, gaze and speech. A number of papers (Fujie et al., 2004; Morency et al., 2009; Morency et al., 2005; Morency et al., 2007; Navarretta and Paggio, 2010) describe classification experiments testing the correlation between speech, prosody and head movements in annotated multimodal corpora. Machine learning algorithms have also been applied to annotations of hand gestures and the co-occurring referring expressions in order to identify gestural features relevant for coreference resolution (Eisenstein and Davis, 2006; Navarretta, 2011).

Moreover, in the present work, we extend the annotations reported in (Lis, 2012b) with two more form attributes (Movement and Direction). These attributes are chosen because they belong to fundamental parameters of gesture form description (Bressem, 2013) and they are associated with motion, so are expected to be of importance considering we study eventualities, especially motion ones.

The paper is organized as follows. In section 2, we shortly present the background for our study, and in section 3 we describe the multimodal corpus and the annotations used in our analyses. In section 4, we present the machine learning experiments, and in section 5 we discuss our results and their implications, and we propose directions for future research.

2 Background

The form of co-verbal, iconic gestures is influenced by, among others, the semantics of the co-occurring speech and by the visually perceivable characteristics of the entity referred to (Kita and Özyürek, 2003; McNeill, 1992). Poggi (2008) has suggested that not only the observable properties of the referent should be taken into consideration but also "the type of semantic entity it constitutes." She has distinguished four such types (Animates, Artifacts, Natural Objects and Eventualities) and proposed that their gestural representation will differ.

Eventualities themselves can still be represented in gesture in various ways, for example from different Viewpoints (McNeill, 1992; McNeill, 2005). In Character Viewpoint gestures (C-vpt), an eventuality is shown from the perspective of the agent, gesturer mimes agent's behavior; in Observer Viewpoint (O-vpt), the narrator sees the eventuality as an observer and in Dual Viewpoint (D-vpt), the gesturer merges the two perspectives. Parrill (2010) has suggested that the choice of Viewpoint is influenced by the eventuality structure. She has proposed that eventualities which have trajectory as the more salient element - elicit O-vpt gesture, while eventualities in which the use of character's hands in accomplishing a task is more prominent - tend to evoke C-vpt gestures.

Other factors suggested to influence eventuality gestures include verb Aspect and Aktionsart. Aspect marks "different ways of viewing the internal temporal constituency of a situation" (Comrie, 1976). The most common distinction is between perfective and imperfective aspect: the former draws focus to the completeness and resultativness of an eventuality, whereas with the latter the eventuality is viewed as ongoing. Duncan (2002) has analyzed the relationship between Aspect of verbs and Handedness in gestures in English and Chinese data. Handedness regards which hand performs the movement and, in case of bi-handed gestures, whether the hands mirror each other. Duncan has found that symmetric bi-handed gestures more often accompany perfective verbs than imperfective ones; the latter mostly co-occur with two handed non-symmetric gestures. Parrill and colleagues (2013) have investigated the rela-

tionship between verbal Aspect and gesture Iteration (repetition of a movement pattern within a gesture). They have found that descriptions in progressive Aspect are more often accompanied by iterated gestures. This is, however, only the case if eventualities are presented to the speakers in that Aspect in the stimuli.

Aktionsart is a notion similar to, but discernible from, Aspect.[2] It concerns Vendler's (1967) distinction between States, Activities, Accomplishments and Achievements, according to differences between the static and dynamic, telic and atelic, durative and punctual. Becker and colleagues (2011) have conducted a qualitative study on Aktionsart and temporal coordination between speech and gesture. They have suggested that gestures affiliated with Achievement and Accomplishment verbs are completed, or repeated, on the goal of the verb, whereas in case of gestures accompanying Activity verbs, the stroke coincides with the verb itself.

Lis (2012a) has introduced a framework in which the relationship between these factors and gestural expressions of eventualities is investigated using wordnet databases, i.e. electronic linguistic taxonomies. She has employed wordnet to, among others, formalize Poggi (2008) and Parrill's (2010) insights. Based on plWordNet 1.5 classification, she has distinguished different types of eventualities and showed their correlation with gestural representation (Lis, 2012b). The present study further builds up on that work, using updated (plWN 2.0), revised (Lis, submitted) and extended annotations and machine learning experiments.

3 The data

3.1 The corpus

Our study was conducted on the refined annotations (Lis, submitted) from the corpus described in (Lis, 2012a; Lis, 2012b), which has in turn been an enriched version of the PCNC corpus created by the DiaGest research group (Karpiński et al., 2008). Data collection followed the well-established methodology of McNeill (1992; 2005): the corpus consists of audio-video recordings of 5 male and 5 female adult native Polish speakers who re-tell a Canary Row cartoon to an addressee. The stimulus contains numerous eventualities and has proved to elicit rich multimodal output. The monologues were recorded in a studio as shown in Figure 1 and the whole corpus consists of approximately one hour of recordings.

[2]For a discussion on the differences between Aspect and Aktionsart and between the Germanic and Slavic traditions of viewing these two concept cf. (Młynarczyk, 2004).

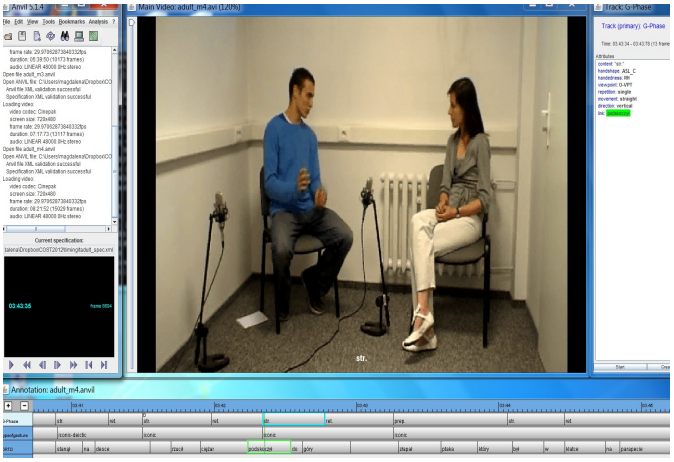

Figure 1: A snapshot from the ANVIL tool

3.2 The annotation

Speech has been transcribed with word time stamps by the DiaGest group, who has also identified communicative hand gestures and annotated their phases, phrases and semiotic types in ELAN (Wittenburg et al., 2006). Lis (2012a; 2012b) exported the annotations to the ANVIL tool (Kipp, 2004) and enriched it with coding of verbs and Viewpoint, Handedness, Handshape and Iteration of gestures. The annotations in the corpus were refined and, for the purpose of the present study, further extended with two more gesture form attributes (Direction and Movement) (Lis, submitted).

3.2.1 The annotation of gestures

Iconic hand gestures were identified based on DiaGest's annotation of semiotic types. Gestures depicting eventualities were manually annotated using six pre-defined features, as reported in detail in (Lis, submitted). Table 1 shows the attributes and values for gestures annotation used in this study. Viewpoint describes the perspective adopted by the speaker and was encoded using the values proposed by McNeill: C-, O- and D-vpt (1992). The attribute Handedness indicates whether one (*Right_Hand*, *Left_Hand*) or two hands are gesturing and whether they are symmetric or not (*Symmetric_Hands* versus *Non-symmetric_Hands*). Handshape refers to configu-

Table 1: Annotations of gestures

Attribute	Value
Viewpoint	Observer_Viewpoint, Character_Viewpoint, Dual_Viewpoint,
Handshape	ASL_C, ASL_G, ASL_5, ASL_O, ASL_S, Complex Other
Handedness	Right_Hand, Left_Hand, Symmetric_Hands, Non-symmetric_ Hands
Iteration	Single, Repeated, Hold
Movement	Straight, Arc, Circle, Complex, None
Direction	Vertical, Horizontal, Multidirectional, None

ration of palm and fingers of the gesturing hand(s); the values are taken from American Sign Language Handshape inventory (Tennant and Brown, 2010): *ASL_C, ASL_G, ASL_5, ASL_O, ASL_S* and supplemented with the value for hand shapes changing throughout the stroke (*Complex*) or not falling under any of the mentioned categories (*Handshape_Other*). Iteration indicates whether a particular movement pattern within a stroke occurs once (*Single*) or multiple times (*Repeated*), or whether the stroke consists of a static *Hold*. Movement regards shape of the motion, while Direction - the plane on which the motion is performed.

3.2.2 The annotation of verbs

Verbs were identified in the word stamp speech transcript. Information about verbs was extracted from the Polish WordNet, plWordNet 2.0, following the procedure explained in (Lis, 2012a; Lis, 2012b). In a wordnet, the lexical units are classified into sets of synonyms, called synsets, which are linked to each other via a number of conceptual-semantic and lexical relations (Fellbaum, 1998). The most frequently encoded one is hyponymy, also called IS_A or TYPE_OF relation, that connects a sub-class to its superclass, the hyperonym. Non-lexical synsets in the upper-level hierarchies of hyponymy encodings in plWordNet contain information on verb Aspect, Aktionsart and domain (Maziarz, 2012).

A domain denotes a segment of reality and all lexical units belonging to a particular domain share a common semantic property (Brinton, 2000). Lis (2012a; 2012b) has used wordnet domains to categorize referents of multimodal expressions according to their type. The attribute Eventuality Type was assigned based the domain of the verb used in speech to denote the eventuality. The choice of the domains in focus has been partially inspired by Parrill's distinction between eventualities with a more prominent trajectory versus eventualities with a more prominent handling element (Parrill, 2010). Based on this, Lis (2012a; 2012b) has distinguished two Eventuality Types:[3] Translocation and Body_Motion. The former refers to eventualities with traversal of a path of a moving object or focus on spatial arrangement and the latter refers to a movement of agent's body (part) not entailing displacement of the agent as a whole (cf: (Levin, 1993)). Lis has subsumed plWordNet domains to fit this distinction. The domains relevant to our study are (with examples of verbs from the corpus given in parentheses):

TRANSLOCATION

{location or spatial relations}[4](*spadać* 'to fall,' *zderzać się* 'to collide');

{change of location or spatial relations change}(*biegać* 'to run,' *skakać* 'to jump').

BODY_MOTION

{causing change of location or causing spatial relations change}(*rzucać* 'to throw,' *otwierać* 'to open');

{physical contact}(*bić* 'to beat', *łapać* 'to catch');

{possession}(*dawać* 'to give,' *brać* 'to take');

{producing}(*budować* 'to build,' *gotować* 'to cook').

Verbs from the synsets {location or spatial relations} and its alterational counterpart were subsumed under the type Translocation. More examples of the verbs from the corpus include: *wspinać się* 'to climb,'*chodzić* 'to walk,' *wypadać* 'to fall out'. Synsets {causing change of location or causing spatial relations change} and {physical contact}, as well as {possession} and {producing} were grouped under the type Body_Motion. Further verb examples are: *przynosić* 'to bring,' *trzymać* 'to keep,' *walić* 'to bang,' *dawać* 'to give,' *szyć* 'to sew.' Verbs from the remaining domains were collected under the umbrella term 'Eventuality_Other.' These verbs constituted less than 10% of all verb-gesture tokens found in the data. Examples include: {social relationships} *grać* 'to play,' {mental or emotional state} *oglądać* 'to watch.' For the purpose of the analyses in the present paper, they were combined with

[3]Note that these categories are orthogonal to Poggi's (2008) ontological types.

[4]In wordnets {} indicates a synset.

Table 2: Annotations of verbs

Attribute	Value
Eventuality Type	Translocation, Body_Motion, Other
Aspect	Perfective,Imperfective
Aktionsart	State, Act, Activity, Accident, Event, Action, Process

Table 3: Classification of Handshape

Hanshape	Precision	Recall	F-score
baseline	0.08	0.28	0.12
Aspect	0.08	0.28	0.12
Aktionsart	0.22	0.28	0.21
Type	**0.17**	**0.32**	**0.22**
all	0.19	0.27	0.21

the Body_Motion category.[5] The domains were semi-automatically assigned to the verbs in our data. Verb polysemy was resolved with a refined version (Lis, submitted) of the heuristics proposed in (Lis, 2012b).

Apart from the domains, the encoding of hyponymy-hyperonymy relations of verbs in plWordNet provides also information about Aktionsart and Aspect. The attribute Aspect has two possible values: Perfective and Imperfective. For Aktionsart, seven categories are distinguished: States, Acts, Activities, Accidents, Events, Actions and Processes. They are Laskowski's (1998) adaptation of Vendler's (1967) Aktionsart classification to the features typical for Polish language.[6] Table 2 shows the attributes and values for verbs annotation used in our study.

3.2.3 The annotation process

Gestures and verbs were coded on separate tracks and connected by means of MultiLink option in ANVIL. Gestures were linked to the semantically affiliated verb. The verbs and gestures were closely related temporally: 80% of the verb onsets fell within stroke phase or slightly preceded it (Lis, submitted). Figure 1 shows a screen-shot of the annotations in the tool. 269 relevant verb-gesture pairs were found in the data. Intercoder agreement was calculated for the majority of the gesture annotation attributes and ranged from 0.67 to 0.96 (Lis, submitted) in terms of κ score (Cohen, 1960), i.e. from substantial to almost perfect agreement (Rietveld and van Hout, 1993).

4 The classification experiments

In the machine learning experiments we wanted to test to which extent we can predict the form of hand gestures from the characteristics of eventualities and verbs, as reflected in plWordNet's categorization. The relevant data were extracted from gesture and orthography tracks in ANVIL, and combined using the Multilink annotation. Classification experiments were performed in WEKA (Witten and Frank, 2005) using ten-fold cross-validation to train and test the classifiers. As baseline in the evaluation, the results obtained by the ZeroR classifier were used. ZeroR always chooses the most frequently occurring nominal value. An implementation of a support vector classifier (WEKA's SMO) was applied in all other cases; various algorithms were tested, with SMO giving the best results. The results of the experiments are provided in terms of Precision, Recall and F-score (Witten and Frank, 2005).

4.1 Classifying the gesture form features from linguistic information

In these experiments we wanted to test whether it is possible to predict the form of the gesture from the type of the eventuality referred to and information about Aspect and Aktionsart. The first group of experiments regards the Handshape attribute with seven possible values. In Table 3, the results of these experiments are shown. They indicate that Aspect information does not at all affect the classification of Handshape, and Eventuality Type and Aktionsart only slightly contribute to the classification (the best result is obtained using Eventuality Type annotation, F-score improvement of 0.1 with respect to the baseline, but is not significant).[7] Not surprisingly, the confusion matrix from this experiment shows that the categories which are assigned more correctly are those that occur more often in the data (*ASL_5* and *ASL_S*).

In the following experiment, we wanted to test whether Aktionsart, Aspect and Eventuality Type are related to the employment of hands in the gestures. Thus, Handedness was predicted using the

[5] The resulting frequency distribution of Type in the verb-gesture pairs: Translocation(150) and Body_Motion+Other(119).

[6] Laskowski's (1998) categories of Vendler's (1967) Aktionsart are called Classes. For the sake of simplicity, we use the term Aktionsart instead of Class to refer to them.

[7] We indicate significant results with $*$. Significance was calculated with one-tailed t-test and $p<0.05$.

Table 4: Classification of Handedness

Handedness	Precision	Recall	F-score
baseline	0.2	0.44	0.27
Aspect	0.2	0.44	0.27
Aktionsart	0.33	0.45	0.37
Type	**0.36**	**0.48**	**0.41**
all	0.35	0.47	0.40

Table 5: Classification of Iteration

Iteration	Precision	Recall	F-score
baseline	0.55	0.74	0.63
Aspect	0.55	0.74	0.63
Aktionsart	0.55	0.74	0.63
Type	0.55	0.74	0.63
all	0.55	0.74	0.63

verb related annotations. The results of these experiments are in Table 4. Also in this case, Aspect does not contribute to the prediction of gesture form. However, the results show that information about the Eventuality Type to some extent improves classification with respect to the baseline (F-score improvement: 0.14∗). The most correctly identified gestures were performed with *Right_Hand* and *Symmetrical_Hands*, which are the most frequently occurring Handedness values in the data.

In the third group of experiments, we wanted to investigate whether the linguistic categorization of verbs improves the prediction of the gesture Iteration. The results of these classification experiments are in Table 5. They indicate that no single feature contributes to the classification of hand repetition: in all cases the most frequently occurring value, *Single*, is chosen as in the baseline.

In the fourth group of experiments we analyzed whether the linguistic categorization of verbs enhances the prediction of Movement. We present the results of these classification experiments in Table 6. They show that none of the investigated verbal attributes has a relation to the Movement in gesture.

In the fifth group of experiments the relation between the linguistic categorization of verbs and the direction of the hand movement was determined. The results of these classification experiments are given in Table 7. They indicate that only Aktionsart contributes to the prediction of Direction (the improvement with respect to the baseline: 0.16∗).

Table 6: Classification of Movement

Movement	Precision	Recall	F-score
baseline	0.37	0.61	0.46
Aspect	0.37	0.61	0.46
Aktionsart	0.37	0.61	0.46
Type	0.37	0.61	0.46
all	0.37	0.61	0.46

Table 7: Classification of Direction

Direction	Precision	Recall	F-score
baseline	0.26	0.50	0.34
Aspect	0.26	0.50	0.34
Aktionsart	**0.47**	**0.55**	**0.50**
Type	0.26	0.50	0.34
all	0.47	0.55	0.50

Table 8: Predicting the Viewpoint type from linguistic information

Viewpoint	Precision	Recall	F-score
baseline	0.29	0.54	0.38
Aspect	0.29	0.54	0.38
Aktionsart	0.53	0.59	0.53
Type	**0.71**	**0.78**	**0.74**
all	**0.71**	**0.78**	**0.74**

4.2 Classifying the Viewpoint

In the following experiments we investigated to what extent it is possible to predict the Viewpoint in gesture from a) the linguistic categorization of the verb and b) from the gesture form.

In the first experiment, we tried to automatically identify the Viewpoint in the gesture from the Eventuality Type annotation. We also investigated to which extent the verb Aspect and Aktionsart contribute to the classification. The results of these experiments are in Table 8. The results confirm that there is a strong correlation between Viewpoint and Eventuality Type (F-score improvement with respect to the baseline: 0.36∗). We also found a correlation between Viewpoint and Aktionsart.

In Figure 2 the confusion matrix for the best classification results are given. Not surprisingly, the classifier did not perform well on the very infrequent category, i.e. D-vpt.

```
  a   b   c   <-- classified as
 89   0  12 |   a = C-VPT
  5   0  18 |   b = D-VPT
 25   0 120 |   c = O-VPT
```

Figure 2: Confusion matrix for predicting Viewpoint from linguistic information

In the last group of experiments we applied the SMO classifier to the data to predict Viewpoint

Table 9: Predicting the Viewpoint type from form features

Viewpoint	Precision	Recall	F-score
baseline	0.29	0.54	0.38
Handshape	**0.64**	**0.7**	**0.67**
Handedness	**0.58**	**0.64**	**0.60**
Iteration	0.67	0.57	0.44
Movement	0.55	0.55	0.43
Direction	0.67	0.57	0.44
all	**0.68**	**0.72**	**0.69**

from Handshape, Handedness, Iteration, Movement and Direction. Table 9 summarizes the results of these experiments. They demonstrate a strong correlation between the form of a gesture and the gesturer's Viewpoint: F-score improvement with respect to the baseline is $0.31*$ when all form related features are used, and all features contribute to the classifications. Handshape and Handedness are the features most strongly correlated to Viewpoint. In Figure 3 the confusion matrix for the best classification results is given.

```
   a   b   c   <-- classified as
84   0  17 |   a = C-VPT
20   0   3 |   b = D-VPT
36   0 109 |   c = O-VPT
```

Figure 3: Confusion matrix for predicting Viewpoint from form features

5 Discussion and future work

The results of our first group of experiments indicate that it is to some extent possible to automatically predict certain form characteristics of hand gestures from the linguistic categorization of their lexical affiliates. We found that the Eventuality Type extracted from wordnet categorization of verbs improves classification of Viewpoint in the co-occurring gesture. Our results are in line with Lis' (2012b) claim that the type of referent influences gestural representation. This claim has in turn been inspired by Poggi (2008) and Parrill's (2010) hypotheses.

Lis (submitted) interprets the finding in terms of Gricean Maxims (Grice, 1976), which among others state that speakers tend to convey as much relevant information in as economic way as possible. Body Motion refers to a movement of agent's body (part) not entailing displacement of the agent as a whole, which can be easily mimed with hand gestures from an internal perspective. The trajectory or spatial arrangement of Translocation eventualities, on the other hand, is less readily reenacted without the risk of hindering communicative flow between interlocutors. It can, however, be easily depicted from an external perspective with gestures drawing paths. Moreover, we have identified the form features of gestures which are most tightly related to the Viewpoint, that is Handshape and Handedness. In line with the previous interpretation, Lis (submitted) suggests that C-vpt gestures often depict interaction with an object and the hand shapes reflect grasping and holding. O-vpt gestures, on the contrary, focus on shapes and spatial extents and utilize, thus, hand shapes convenient for depicting lines, i.e. a hand with extended finger(s). It needs to be, however, further examined in how far the distribution of Handshape and Handedness in our data is motivated by the specifics of the stimuli.

Our findings also show that the type of eventuality improves prediction of Handedness. However, Eventuality Type provides a more substantial improvement in the prediction of Viewpoint, i.e. aspect of gestural representation rather than of purely physical form of gesture. This suggests that considering such representational format as an intermediate step in modeling gesture production may be appropriate. Having found that referent properties are only partially predictive of the form of iconic gesture, Kopp and colleagues (2008) consider direct meaning-form mapping to have a weak empirical support. They have instead suggested a two-step micro-planning procedure where the relationship between referent properties and gesture physical form is mediated by representational format. The present experiments do not provide an answer as to whether the two-step approach could lead to modeling aspects of eventuality gesture production. More analyses are needed, and they should be addressed in future work.

While our results indicate that Eventuality Type is the strongest predictor of gesture form, we have also found that Handedness and Viewpoint are related to Aktionsart, whereas none of the considered form features showed correlation with verb Aspect. An explanation might be that both the Eventuality Type and Aktionsart regard more inherent characteristics of eventuality, while Aspect regards the speaker's external perspective on the eventuality. It also needs to be noted that not all Aktionsart categories are equally represented in

our data.[8] The three most frequent Aktionsart categories share the feature 'intentionality,' but belong to different groups in Vendler's classification (Maziarz et al., 2011). It should be investigated in how far different Aktionsart types in our data are represented for different Eventuality Types, as that may provide a further explanation of the obtained results.

Aspect does not improve the classification for any feature. The observation that Aspect is related to Handedness (Duncan, 2002) and Iteration (Parrill et al., 2013) is, thus, not reflected in this corpus. It needs to be remembered that the relationship between Aspect and Iteration was found by Parrill and colleagues (2013) only when the eventualities were presented to speakers in the appropriate Aspect in the stimuli. Our results suggest it may not be generalizable to an overall correlation between Aspect and gesture Iteration. Moreover, Aspect is expressed very differently in the three languages under consideration (Polish - the present study, English (Parrill et al., 2013), and English and Chinese (Duncan, 2002)). Cross-linguistic differences have been found to be reflected in gesturing (Kita and Özyürek, 2003). Whether such differences in encoding of Aspect impact gestures should be, thus, investigated further.

The results of the experiments also indicate that gestural Iteration and Movement are not at all related to the linguistic characteristics of the co-occurring verb and that the only feature improving classification of gesture direction is Aktionsart. For Iteration, however, our data are biased in that single gestures are predominant, which may have affected the results. Regarding Movement and Direction, we suggest that they may be primarily dependent on visual properties of the referent, rather than the investigated factors. For example, Kita and Özyürek (2003) have found that the direction of gesture in elicited narrations reflects the direction in which an eventuality has been presented in the stimuli. The only improvement identified in our experiments in the classification of Direction (due to Aktionsart) requires further investigation.

Our results suggest the viability of the framework adopted in the paper, i.e. application of wordnet for investigation of speech-gesture ensembles. Wordnet classification of lexical items can be used to shed some light on speech-related gestural behavior. Using wordnet as an external source of annotation increases coding reliability and due to the wordnet machine-readable format, it enables automatic assignment of values. Wordnets exist for numerous languages and the approach may, thus, be applied cross-linguistically and help to uncover universal versus language-specific structures in gesture production. The findings support the viability of a number of categories in the annotation scheme used - they corroborate that the type of referent is a category relevant to studying gestural characteristics and they validate the importance of introducing distinctions among eventualities for multimodal phenomena. The experiments also identify another attribute, i.e. Aktionsart, as relevant in the framework.

It has to, however, be noted that our study is only preliminary, because the results of our machine learning experiments are biased by the fact that for some attributes certain values occur much more frequently than others in the data. Future work should address normalization as a possible solution. Moreover, our findings are based on narrational data, and need to be tested on different types of interaction. Most importantly, the dataset we used is small for machine learning purposes. Due to time load of multimodal annotation, small datasets are a well-known challenge in gesture research. Our results await, thus, validation on a larger sample. Also, cross-linguistic studies on comparative corpora should be performed.

In the present work only one type of bodily behaviors, i.e. hand gestures, was taken into account, but people use all their body when communicating. Thus, we plan to extend our investigation to gestures of other articulators, such as head movements and posture changes. In the present work only gestures referring to eventualities were considered. Lis (submitted) has recently started extending the wordnet-based framework and investigation to animate and inanimate objects.

[8] The frequency distribution of Aktionsart in the verb-gesture pairs: Activities(115), Acts(56), Actions(58), Events(23), Accidents(15), States(2), Processes(0), and of Aspect: Imperfective(179) and Perfective(126).

References

Polish WordNet. Wrocław University of Technology. http://plwordnet.pwr.wroc.pl/wordnet/.

Tennant, R. and M. Brown *The American Sign Language Handshape Dictionary.* Washington, DC: Gallaudet University Press (2010).

Alibali, M. W., Heath, D. C., and Meyers, H. J. Effects of visibility between speakers and listeners on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44:159–188 (2001).

Becker, R., Cienki, A., Bennett, A., Cudina, C., Debras, C, Fleischer, Z., M. Haaheim, T. Mueller, K. Stec, and A. Zarcone. Aktionsarten, speech and gesture. In *Gesture and Speech in Interaction '11*, (2011).

Bressem, J. A linguistic perspective on the notation of form features in gestures. *Body – Language – Communication. Handbooks of Linguistics and Communication Science.* Berlin, New York: Mouton de Gruyter (2013).

Brinton, L. *The structure of modern English: A linguistic introduction.* John Benjamins Publishing Company (2000).

Comrie, B. *Aspect.* Cambridge: Cambridge University Press (1976).

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46 (1960).

Duncan, S. Gesture, verb Aspect, and the nature of iconic imagery in natural discourse. *Gesture*, 2(2):183–206 (2002).

Eisenstein, J.and Davis, R. Gesture features for coreference resolution. In Renals, S., Bengio, S., and Fiscus, J., editors, *MLMI 06*, pages 154–155 (2006).

Eisenstein, J.and Davis, R. Gesture improve coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 37–40, New York (2006).

Fellbaum, C., *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA (1998).

Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y., and Kobayashi, T. A conversation robot using head gesture recognition as para-linguistic information. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*, 159–154 (2004).

Grice, H. *Logic and Conversation. Syntax and Semantics*, 3:41–58. Academic Press, New York (1976).

Jokinen, K. , Navarretta, C., and Paggio, P. Distinguishing the communicative function of gesture. *Proceedings of MLMI* (2008).

Karpiński, M., Jarmołowicz-Nowikow, E., Malisz, Z., Szczyszek, M., Juszczyk, J. Rejestracja, transkrypcja i tagowanie mowy oraz gestów w narracji dzieci i dorosłych. *Investigationes Linguisticae*, 17 (2008).

Kendon, A. *Gesture: Visible Action As Utterance*. Cambridge University Press, Cambridge (2004).

Kipp, M. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Boca Raton, Florida (2004).

Kita, S. and A. Özyürek. What does cross–linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1):16–32 (2003).

Kopp, S., Bergmann, K., and Ipke, W. Multimodal communication from multimodal thinking – towards an integrated model of speech and gesture production. *Semantic Computing*, 2(1):115–136 (2008).

Krauss, R. M., Chen, Y., and Gottesman, R. F. Lexical gestures and lexical access. a process model. In McNeill, D., editor, *Language and Gesture*, pages 261–283. Cambridge University Press, New York (2000).

Laskowski, L. *Kategorie morfologiczne języka polskiego — charakterystyka funkcjonalna.* PWN, Warszawa (1998).

Levin, B. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago (1993).

Lis, M. Annotation scheme for multimodal communication: Employing plWordNet 1.5. In *Proceedings of the Formal and Computational Approaches to Multimodal Communication Workshop. 24th European Summer School in Logic, Language and Information (ESSLLI'12)* (2012).

Lis, M. Influencing gestural representation of eventualities: insights from ontology. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI'12)*, 281–288, (2012).

Lis, M. Multimodal representation of entities: A corpus-based investigation of co-speech hand gesture. PhD dissertation, University of Copenhagen (submitted).

Louwerse, M., Jeuniaux, P., Hoque, M., Wu, J., and Lewis, G. Multimodal communication in computer-mediated map task scenarios. In Sun, R. and Miyake, N., editors, *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Erlbaum (2006).

Louwerse, M. M., Benesh, N., Hoque, M., Jeuniaux, P., Lewis, G., Wu, J., and Zirnstein, M. Multimodal communication in face-to-face conversations. In Sun, R. and Miyake, N., editors, *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Erlbaum (2006).

Maziarz, M. Non-lexical verb synsets in upper-hierarchy levels of polish wordnet 2.0. Technical report, Wrocław University of Technology (2012).

Maziarz, M., Piasecki, M., Szpakowicz, S., Rabiega-Wiśniewska, J. and B. Hojka. Semantic relations between verbs in polish wordnet 2.0. *Cognitive Studies*, (11):183–200 (2011).

McNeill, D. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago (1992).

McNeill, D. *Gesture and Thought.* University of Chicago Press, Chicago (2005).

Melinger, A. and Levelt, W. Gesture and the communicative intention of the speaker. *Gesture*, 4(2):119–141 (2005).

Młynarczyk, A. Aspectual pairing in Polish. PhD dissertation, University of Utrecht (2004).

Morency, L.-P., de Kok, I., and Gratch, J. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70–84 (2009).

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. Contextual recognition of head gestures. In *Proceedings of the International Conference on Multimodal Interfaces* (2005).

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171(8–9):568–585 (2007).

Navarretta, C. Anaphora and gestures in multimodal communication. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 171–181, Faro, Portugal (2011).

Navarretta, C. and Paggio, P. Classification of feedback expressions in multimodal data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 318–324, Uppsala, Sweden (2010).

Parrill, F. Viewpoint in speech–gesture integration: Linguistic structure, discourse structure, and event structure. *Language and Cognitive Processes*, 25(5):650–668 (2010).

Parrill, F., Bergen, B. and P. Lichtenstein. Grammatical aspect, gesture, and conceptualization: Using co-speech gesture to reveal event representations. In *Cognitive Linguistics*, 24(1): 135–158 (2013).

Peirce, C. S. *Collected Papers of Charles Sanders Peirce (1931-58)*. Hartshorne, P. Weiss and A. Burks, Cambridge, MA: Harvard University Press (1931).

Poggi, I. Iconicity in different types of gestures. *Gesture*, 8(1):45–61 (2008).

Ramchard, G. Post-davidsionianism. *Theoretical Linguistics*, 31(3):359–373 (2005).

Rauscher, F. H., Krauss, R. M., and Chen, Y. Gesture, Speech, and lexical access: The Role of Lexical Movements in Speech Production. *Psychological Science*, 7(4):226–231 (1996).

Rietveld, T. and Hout, R. v. *Statistical Techniques for the Study of Language and Language Behavior*. Mouton De Gruyter, Berlin (1993).

Vendler, Z. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY (1967).

Witten, J. and Frank, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2 edition (2005).

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. Elan: a professional framework for multimodality research. In *LREC'06, Fifth International Conference on Language Resources and Evaluation* (2006).